Automated Scoring of Student Essays for Diagnostic Assessments

Jason M. Bryer[1], Angela L. Lui[1], Anthony Fraser[1], & Julia Ferris[1]

[1] City University of New York - School of Professional Studies

Author note

DAACS was developed under grants #P116F150077 and #R305A210269 from the U.S. Department of Education. However, the contents do not necessarily represent the policy of the U.S. Department of Education, and you should not assume endorsement by the Federal Government.

Correspondence concerning this article should be addressed to Jason M. Bryer, 119 W 31st St, New York, NY 10001. E-mail: jason.bryer@cuny.edu

Abstract

Significant advancements in large language models have occurred over the past decade. This study explores how some of the more recent tokenizers compare to traditional *n*-gram-style tokenization procedures in the context of scoring a diagnostic assessment. Results show improvements, but predictive models might matter more.

Keywords: natural language processing, machine learning, automated essay scoring

Automated Scoring of Student Essays for Diagnostic Assessments

The use of automated scoring systems has been an area of great interest for educators, especially for written work. Significant advancements have occurred in large language models (LLMs) over the past decade, leading to continuous and rapid growth in automated scoring systems. In this proposal, we compare the performance of automated scoring using recently developed LLMs to one developed a decade ago in the context of DAACS.

Briefly, DAACS, short for Diagnostic Assessment and Achievement of College Skills (Authors, in press), is a free, open-source suite of diagnostic assessments designed to assess college readiness in self-regulated learning (SRL), writing, reading, and mathematics, and provide immediate personalized feedback, including recommended strategies and resources to optimize student learning. The writing assessment, which is the focus of this proposal, asks students to write a 350-word essay reflecting on their SRL results and feedback (Appendix A) and is assessed on nine criteria (see Appendix B).

Theoretical and Contextual Background

Essay scoring is time- and resource-intensive, especially with human scorers, hence the necessity to adopt an automated scoring system for DAACS. In 2014, the Hewlett Foundation sponsored a Kaggle competition (Hamner, 2012) that resulted in approaches with reliability estimates comparable to those of human-to-human scores (Shermis, 2014). One open-source project that emerged from this competition was LightSide (Mayfield, Adamson & Rosé, 2015), the scoring algorithm for the writing component of DAACS.

Automated scoring systems typically have two phases:

1. *Tokenization*: a corpus of text/essays is converted into a numeric matrix.

2. *Modeling*: predictive models (e.g. logistic regression, classification trees) are trained and used to predict scores for new essays.

LightSide and other automated scoring algorithms available in the early 2010s largely relied on variations of encoding text into *n*-grams with pre-processing steps (e.g., stemming). However, significant advances in LLP have occurred in the last decade. In 2018, Google introduced the Bidirectional Encoder Representations from Transforms (BERT) language model (Devlin, Chang, Lee, & Toutanova, 2018), a self-supervised learning model in which a text is represented as a sequence of vectors.

This project aims to upgrade the current version of the DAACS writing assessment, which uses LighSside, a decade-old LLM, to a more cutting-edge approach. For comparison, Table 1 provides the human-to-human and LightSide-to-human inter-rater reliability across the nine domains.

Table 1

Inter-rater reliability of human raters and accuracy of LightSide.

Criteria	Human-Human	Model	Human-LightSide
Conntent Suggestions	0.61	Logit	0.72
Content Summary	0.56	Logit	0.70
Conventions	0.55	Logit	0.63
Organization Structure	0.63	Bayes	0.74
Organization Transitions	0.57	Bayes	0.47
Paragraphs Cohesion	0.62	Logit	0.73

Criteria	Human-Human	Model	Human-LightSide
Paragraphs Focus on a Main Ideas	0.60	Logit	0.73
Sentences Complexity	0.56	Bayes	0.68
Sentences Correct	0.56	Logit	0.56

Methods

Data Source

The data for this study were drawn from a larger dataset collected in 2016 as part of a large-scale randomized control trial to test the efficacy of DAACS. Student demographics are provided in Appendix C.

Essays were scored by a team of 12 expert raters and 413 (46.20%) were scored by two raters. Raters conferred on any discrepancies and agreed upon a "true" score. For essays that were double scored the conferred score was used for training, otherwise the single rater scores were used.

Procedures

R and Python were used to explore six tokenization algorithms (Table 2) across nine predictive models (Table 3) for the nine criteria for the DAACS writing assessment. All essays were tokenized using the six methods in Table 2. Essays were then were randomly divided into a training (70% of essays) and validation (30% of essays) data sets. Predictive models were trained using each of the nine models listed in Table 3 across the nine criteria and accuracy recorded using the validation data set resulting in 486 models.

Table 2

Tokenization algorithms

Tokenization Algorithm	Reference
Bert	Devlin et al. (2018)
Distilbert	Sanh, Debut, Chaumond, and Wolf (2019)
Word 2 Vec	Mikolov, Chen, Corrado, and Dean (2013)
Scikit Native Vectors	Arsenovic et al. (2022)
Facebook Vectors	Bojanowski et al. (2016)
Scikit Vectors	Arsenovic et al. (2022)

Table 3

Predictive models

Prediction Model	Reference
k-nearest neighbors	Fix & Hodges (1951)
Linear Support Vector Classification	Ratna et al (2019)
C-Support Vector Classification	Novakovic et al (2011)
Gradient Boosting	Breiman (1997)
XGBoost	Shang, Men, and Du (2023)
Bagging	Islam et al (2022)

Prediction Model	Reference
Random Forest	Breimann (2001)
AdaBoost	Sevinç (2022)
Logistic Regression	Cramer (2002)

Results

Figure 1 plots the model accuracy for all combinations of tokenizers and models. Table 4 provides the combination with the highest predictive accuracy for each criterion. Regarding tokenizers, the BERT class had the highest performance in all but one criterion; regarding models, random forests had the highest performance in six of the nine criteria. Further examination of Figure 1 reveals that random forests (and also gradient boosting and bagging) had very small differences in accuracy among different tokenizers, suggesting that random forests perform very well regardless of the tokenizers used.

Figure 1

Prediction accuracy for the nine criteria grouped by predictive model (top) and tokenizer (bottom).

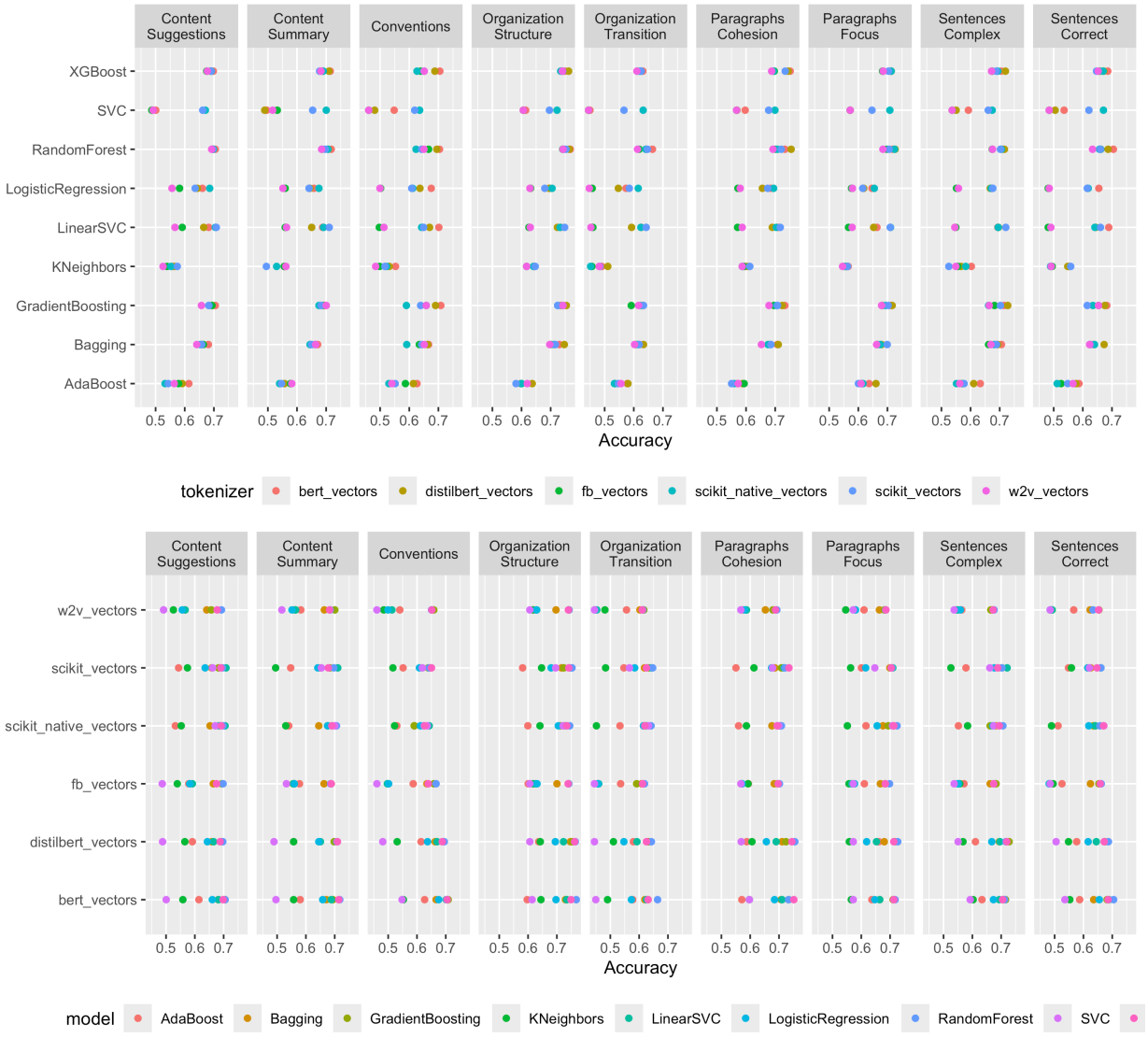

Table 4

Best prediction methods for each domain.

Criteria	Toeknizer	Predictive Model	Accuracy
Content Suggestions	scikit_vectors	LinearSVC	0.71
Content Summary	bert_vectors	RandomForest	0.72
Conventions	bert_vectors	GradientBoosting	0.71

Criteria	Toeknizer	Predictive Model	Accuracy
Organization Structure	bert_vectors	RandomForest	0.77
Organization Transition	bert_vectors	RandomForest	0.66
Paragraphs Cohesion	distilbert_vectors	RandomForest	0.75
Paragraphs Focus on a Main Idea	distilbert_vectors	RandomForest	0.73
Sentences Complex	distilbert_vectors	GradientBoosting	0.73
Sentences Correct	bert_vectors	RandomForest	0.70

Discussion

This study demonstrates improvements in automated scoring systems over the decade. In all but one criterion, the newer tokenizers increased the prediction accuracy by 2% to 19%. It should be noted that in all cases, the accuracy of the automated scoring system was higher than the human-to-human accuracy. This discrepancy needs to be further explored, especially with regard to high-stakes assessments; perhaps the difference can be explained by the fact that computers do not experience fatigue. Nonetheless, this is promising for using automated scoring for formative assessment where timely feedback is critical.

Additionally, this study suggests that the choice of predictive models may be more important than the choice of tokenizers. The environmental impact of training large language models such as BERT has been well documented (Bender, Gebru, McMillan-Major, & Shmitchell, 2021). At least in the context of automated scoring, using less computationally intensive tokenizers might provide comparable results.

References

Arsenovic, A., Hillairet, J., Anderson, J., Forstén, H., Rieß, V., Eller, M., … Forstmayr, F. (2022). Scikit-rf: An open source python package for microwave network creation, analysis, and calibration [speaker's corner]. *IEEE Microwave Magazine*, *23*(1), 98–105. https://doi.org/10.1109/MMM.2021.3117139

Ben Hamner, lynnvandev, Jaison Morgan. (2012). *The hewlett foundation: Automated essay scoring*. Kaggle. Retrieved from https://kaggle.com/competitions/asap-aes

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., Douze, M., & Jegou, H. (2016). FastText.zip: Compressing text classification models. *arXiv Preprint arXiv:1612.03651*.

Breiman, L. (2001). *Random forests*. *45*(1). https://doi.org/10.1023/A:1010933404324

Cramer, J. S. (2003). The origins of logistic regression. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.360300

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv. https://doi.org/10.48550/ARXIV.1810.04805

Fix, E., & Hodges, J. L. (1951). *Discriminatory analysis: Nonparametric discrimination: Consistency properties*. USAF School of Aviation Medicine. Retrieved from https://books.google.com/books?id=4XwytAEACAAJ

Islam, P., Khosla, S., Lok, A., & Saxena, M. (2022). *Analyzing bagging methods for language models*. Retrieved from https://arxiv.org/abs/2207.09099

Mayfield, E., Adamson, D., & Risé, C. P. (2015). *LightSide*. Retrieved from http://ankara.lti.cs.cmu.edu/side/download.html

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv. https://doi.org/10.48550/ARXIV.1301.3781

Novakovic, J., & Veljovic, A. (2011). C-support vector classification: Selection of kernel and parameters in medical diagnosis. *SISY 2011 - 9th International Symposium on Intelligent Systems and Informatics, Proceedings*. https://doi.org/10.1109/SISY.2011.6034373

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv. https://doi.org/10.48550/ARXIV.1910.01108

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. *Assessing Writing*, *20*, 53–76. https://doi.org/https://doi.org/10.1016/j.asw.2013.04.001

Appendix A: Writing Prompt

You received information about your learning skills after you took the self-regulated learning (SRL) survey, as well as suggestions for becoming a more effective and efficient learner. Now, in order to reflect on your learning skills and receive feedback on your writing, please use the results from your SRL survey to do your best writing in a brief essay that answers the questions below.

You will need to refer to your SRL survey results and feedback in your essay. We recommend reviewing them, taking notes, and then returning here to write.

Click here to open your SRL results in a new window.

Essays must be at least 350 words in order to be meaningfully scored. Please aim to write a complete, well-developed essay in order to get accurate feedback about how ready you are for academic writing, and what you can do to strengthen your writing skills.

What do your self-regulated learning survey results and the feedback tell you about your learning skills? Use results from the survey and the feedback to support your analysis.

Which suggested strategies from the feedback are you committed to using this term? Explain why you are committed to using those strategies.

Click on Help, then Rubric to review the criteria.

Once you complete the writing assessment, you will be provided feedback and suggestions for resources to use to boost your writing skills, as needed. It is really important that you look at the feedback, because research shows that students who read it and use the resources are more successful than those who simply complete the assessments without viewing the feedback.

Appendix B: Scoring Rubric

Criteria	Mastering	Emerging	Developing
content summary	The essay uses relevant survey results and feedback to provide a detailed summary of the student's strengths and weaknesses in terms of self-regulated learning.	The essay refers to survey results and feedback in the summary of strengths and weaknesses related to self-regulated learning. The summary is a bit thin though: Additional, relevant details and examples would have strengthened it.	The summary of strengths and weaknesses in SRL is under-developed. It might need more detail about the results and feedback, and/or should address both strengths and weaknesses, rather than one or the other.
content suggestions	The discussion of suggestions for improvement in SRL are logically and explicitly related to the survey results and feedback, and developed in sufficient depth.	The essay includes a clear discussion of improvements to be made to SRL, but the connections to the survey results and feedback are not always explicit and clear.	The discussion of possible improvements in SRL is vague and/or disconnected from the strengths and weaknesses discussed in the summary. It might have been stronger if the suggestions for improvement were more specific, and explicitly tied to the results and feedback.
organization structure	The essay is well-organized, with an order and structure that present the discussion in a clear, logical manner.	The essay has a basic structure but does not offer a clear, overall organization that enables a reader to understand the progression of one idea to another. It might have been strengthened by the inclusion of an introduction, topic sentences, conclusion, etc.	The essay is not carefully structured. It might move almost randomly from one point to the next. It could have been strengthened by the use of an introduction, topic sentences, a conclusion, and other organizational features.
organization transitions	Transitions between paragraphs are appropriate and	Transitions between paragraphs include language that signals a	Transitions between paragraphs are missing or ineffective. As a result, the

Criteria	Mastering	Emerging	Developing
	effective, and strengthen the progression of the essay.	shift from one main idea to the next, as needed (e.g. "In terms of monitoring what I do…" "In the mindset category . . ."). The use of additional transitions (e.g. similarly, in addition, next, as a result) would have helped link the paragraphs and strengthen the overall organization of the essay.	paragraphs abruptly shift from one idea to the next. Transitions (e.g. similarly, in addition, next, as a result) would have helped link the paragraphs and strengthen the overall organization of the essay.
paragraphs ideas	Paragraphs are consistently and clearly focused on a main idea or point.	In general, each paragraph has a clear focus. Some might have been strengthened by a topic sentence that establishes the focus, but the main ideas are apparent.	Most or all paragraphs lack a clear focus on a main point or topic, perhaps because too many ideas are packed into one paragraph. Each topic should be fully developed in separate paragraphs. Each paragraph should include a statement of the main point, and every sentence in the paragraph should relate to that point.
paragraphs cohesion	Within paragraphs, the individual sentences are seamlessly linked together; the reader can see the relationship between the ideas or information in one sentence and those in another sentence. The writing explicitly links sentences and ideas using adverbs (e.g., similarly, also, therefore), relative pronouns (e.g., who,	The ideas or information in each sentence within a paragraph are loosely linked together. The connections between sentences would have been clarified by additional or better choices of linking words and phrases, such as adverbs (e.g., similarly, also, therefore), relative pronouns (e.g., who, that, which),	Because the ideas or information in each sentence within a paragraph are not linked together, it is hard for a reader to see the relationship between sentences. Ideas could be connected with adverbs (e.g., similarly, also, therefore), relative pronouns (e.g., who, that, which), conjunctions (e.g., and, or, while, whereas), and/or the repetition of key words, as appropriate.

Criteria	Mastering	Emerging	Developing
	that, which), conjunctions (e.g., and, or, while, whereas), and/or the repetition of key words, as appropriate.	conjunctions (e.g., and, or, while, whereas), and/or the repetition of key words, as appropriate.	
sentences correct	Sentences are correct: no run-ons, fragments, or errors in subject-verb agreement.	The sentences are generally correct. Minor issues with grammar, such as errors in subject-verb agreement (e.g., "the survey results suggests…") do not interfere with meaning.	Problems with grammar make the essay hard to read and understand. More attention to the parts of speech (e.g., nouns, pronouns, verbs, adjectives, etc.) was needed.
sentences complexity	Consistent and appropriate use of a variety of sentence structures, including some sophisticated, complex sentence structures and syntactic forms.	Complex syntactic structures are present but not consistently used; sentence structure is varied but not often sophisticated.	The sentences lack syntactic complexity and vary little, if at all, in structure. The sentences are generally simple in structure (subject-verb-object).
conventions conventions	Spelling, punctuation, and capitalization are correct to the extent that almost no editing is needed.	Spelling, punctuation, and capitalization are correct to the extent that only light editing is needed.	Careful editing was needed to correct frequent, distracting errors in spelling, punctuation, and/or capitalization.

Appendix C: Student Demographics

Characteristic	N = 812[1]
Age	33 (27, 39)
Unknown	76
MilitaryStudent	105 (14%)
Unknown	76
FIRST_GEN_STUDENT	301 (41%)
Unknown	76
CITIZENSHIP_STATUS	
Non-Citizen	10 (1.4%)
Non-Resident Alien	8 (1.1%)
U.S. Citizen	705 (98%)
Unknown	89
GENDER	
Female	393 (54%)
Male	341 (46%)
Unknown	78
ETHNICITY2	
Black	77 (11%)
Am. Indian or Alaskan Native	8 (1.1%)
Multiple	28 (3.9%)
Asian	22 (3.1%)
Hispanic	25 (3.5%)
Native Hawaiian	5 (0.7%)
White	549 (77%)
Unknown	98
HOUSEHOLD_INCOME	
<16k	58 (8.5%)

Characteristic	N = 812[1]
16-25k	71 (10%)
25-35k	100 (15%)
35-45k	96 (14%)
45-65k	120 (18%)
>65k	235 (35%)
Unknown	132
TRANSFER_CREDITS	32 (17, 46)
Unknown	76

[1]Median (IQR); n (%)